Machine Learning Optimization Using Feature Selection for Botnet and Brute Force Attacks Detection in Network Systems

Andri Saputra[1*], Samsudiat[1], Hartanto Kurniawan[1], Cahyono Nugroho[1], Yahya Muhammad[2], Fauzan Adzima Hawari[2], Suryadi[3], Abu Saad Ansari[2], and Nurul Taufiqu Rochman[2,4]

[1] Research Center for Artificial Intelligence and Cybersecurity, National Research and Innovation Agency, South Tangerang, Indonesia
[2] Center of Excellence Applied Nanotechnology, Nano Center Indonesia, South Tangerang, Indonesia
[3] Research Center for Photonics, National Research and Innovation Agency, South Tangerang, Indonesia
[4] Research Center for Advanced Material, National Research and Innovation Agency, South Tangerang, Indonesia
*Corresponding author: andri.saputra@brin.go.id

Abstract. The Intrusion Detection System (IDS) plays a critical role in network systems against cyber threats, in which botnet and brute force are the most identified attacks. Anomaly-based IDS as one detection type of IDS is needed to improve its ability to identify cyber threat characteristics based on machine learning. This paper explores an optimized machine learning approach by combining feature selection techniques, namely the Low Variance Filter and the Pearson Correlation Filter. The benchmark dataset, CICIDS2017, is used to evaluate the model by the Decision Tree algorithm. The results show that the model successfully optimizes cyber threat identification by reducing the number of 83 features to 10 for botnet with 99.5% accuracy and 3s computation time and 15 for brute force with 99.8% accuracy and 4s computation time.

Keywords: IDS, Low Variance Filter, Pearson Correlation Filter, CICIDS2017

1. Introduction

Botnet becomes the most serious threat in cybersecurity [1]. Botnet is one of malicious software which can be remotely controlled through a network system and exploit the information system target. Beside botnets, brute force is also the top cybersecurity threats which can be cracking the credential randomly and automatically. Brute Force attacks is an algorithm Brute will place and find all of the character and length of the password's possibility with a lot of password's combination. The Brute Force algorithm represents a computational approach that systematically explores all possible password combinations by exhaustively testing a vast number of character inputs and password lengths, thus providing a comprehensive solution to password cracking problems. Web Attack Brute Force and SSH - Patator are two types of brute force attacks that are the focus of discussion. One program for monitoring and detecting attacks that occur on the network as well as providing alerts on the type of attacks that are occurring, identifying network patterns, and knowing whether they are normal or dangerous is the Intrusion Detection System (IDS). IDS fundamentally represents a system capable of real-time analysis of data packets, recording all network activities, and actively preventing attacks and misuse.

Anomaly-based Intrusion Detection Systems (IDS), commonly referred to as AIDS, can detect unknown malware and attacks by conducting a comprehensive analysis of the transmitted data. However, the existing literature lacks a standardized benchmarking methodology specifically designed for AIDSs. Our review of the literature pertaining to AIDS reveals several issues in the related research, such as the arbitrary selection of algorithms, parameters, and testing criteria, the utilization of outdated datasets, and insufficient depth in the analysis and validation of the obtained

results. The feature selection technique, which combines the low variance filter and Pearson correlation, is an example of a filter-based feature selection approach proposed by Sheena et al. [9] that aligns with the hypothesis. While the low variance method has been long-established, its exploration has been limited, whereas the Pearson Correlation method is frequently employed in expert systems or correlation measurement techniques. Decision tree is a prediction model that utilizes a tree-like or hierarchical structure, where it transforms data into decision trees and decision rules [1]. The main benefit of using decision trees is their ability to simplify complex decision-making processes, resulting in more interpretable solutions to problems.

2. Previous Research

The research by Aksu, D. [2] utilized three machine learning algorithms, namely Decision Tree, K-Nearest Neighbor, and Support Vector Machine, to classify DDoS attacks on the CICIDS2017 dataset. In this study, 80 irrelevant features were removed, resulting in 30 selected features. The research reported a 99% accuracy for the Decision Tree algorithm, but precision and recall metrics were not mentioned. Kundu et al. [5] conducted research using Information Gain feature selection and correlation techniques to eliminate irrelevant features in the dataset. The study demonstrated that the C4.5 algorithm achieved the highest accuracy of 99.68% using only 17 selected features. Oreski [6] focused on data cleaning and feature selection in the dataset. The results showed a significant reduction in preprocessing time, ranging from 60% to 95% of the total processing time. Saputra [7] proposed a new method to enhance the performance of IDS (Intrusion Detection System) in detecting Bots in the CICIDS2017 dataset. The proposed method combined two statistical feature selection techniques: low variance filter and Pearson correlation. The research resulted in the selection of 15 features from the initial 77 features. Although there was a decrease in the number of features, there was no significant change in accuracy. However, the computation time was reduced from 71 seconds to 5.6 seconds.

3. Methodology

3.1. Experimental Flow

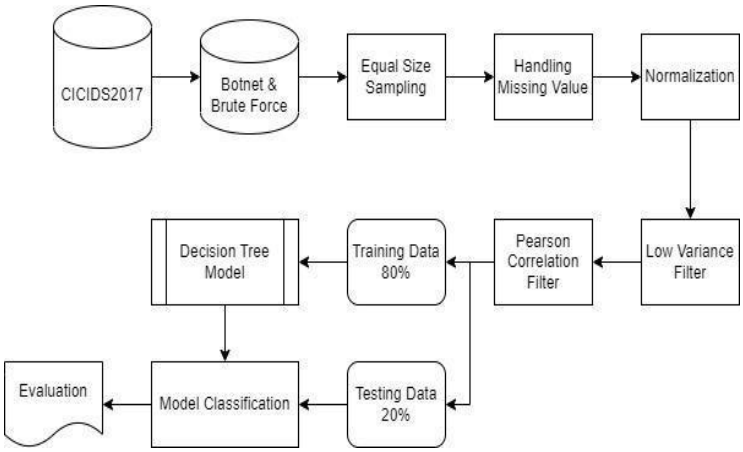

Figure 1. Flow Chart

This research focuses on the use of feature selection methods and building a classification model for Bot and Brute Force attacks using the Decision Tree machine learning algorithm. By employing feature selection methods in developing the training model for machine learning algorithms, the accuracy and computation time of the classification can be determined, and an analysis of the influence of threshold determination on the performance of the model is conducted. The figure illustrates the stages of creating the classification model for Bot and Brute Force attacks.

3.2. Network Traffic Sensor

The used benchmark network traffic sensing dataset, CIC-IDS2017, is published by Canadian Institute for Cybersecurity (CIC), University of New Brunswick (UNB). This dataset contains the network traffic using CICFlowMeter, which can generate bidirectional flows, the forward and backward directions for anomaly detection

[4]. This generated flow is captured with a network traffic sensor and analyzer along with 80 network traffic features such as Destination Port, Flow Duration, Forward Packet Length Max, and Backward Packet Length Max. The CIC-IDS2017 includes 7 scenarios, namely Botnet, Brute Force, DoS/DDoS, Heartbleed, Web Attack, Infiltration, and Port Scan.

3.3. Input Data

The equal size sampling method aims to balance the data taken from CICIDS2017 using Equal Size Sampling. The results are as follows: 1966 data points for the BENIGN class and 1966 for Bot, 1507 data points for the BENIGN class and 1507 for Web Attack Brute Force, and 5897 data points for the BENIGN class and 5897 for SSH - Patator. After that, data cleaning is performed by removing strings and handling missing data. Normalization using the Min-Max method produces data within the range of 0 to 1, which will be further processed in the pre-processing stage. In the pre-processing stage, two methods of feature selection are conducted. The main objective of feature selection is to reduce complexity, improve accuracy, and select the optimal features from a set of data features.

3.4. Selection Feature

Feature selection is an important part of optimizing the performance of classification methods. The main objective of feature selection is to reduce complexity, improve accuracy, and select the optimal features from a set of data features [8]. According to Amiri [3], the filter method commonly used in feature selection is the Pearson correlation coefficient. After the normalization process, the Bot and Brute Force data undergo feature selection. This research utilizes two feature selection methods, namely low variance filter and Pearson correlation exclude. The low variance filter method is calculated using Equation 1, and the Pearson correlation is calculated using Equation 2.

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \tag{1}$$

σ^2 = Variance of Population
X = Feature Value
μ = Mean
N = Total Data

Based on Equation 1, each feature of the attacks undergoes a low variance filter using a specific threshold value [10]. Any attack feature that does not exhibit a strong level of correlation will be removed, in line with the statistical distribution results of the CICIDS2017 dataset, which show a tendency towards a power-law distribution. Power law refers to a condition where minority features can have a significant impact in identifying a class.

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}} \tag{2}$$

r = Pearson Correlation Coefficient
n = Total Data X and Y
Σx = Total of X Variable
Σy = Total of Y Variable

Pearson correlation exclude is a formula used to measure the relationship between features, where in selecting the feature relationships, the individual correlations between them do not affect each other but are independent based on the dominant feature values of each class.

3.5. Data Distribution

The filtered data, after undergoing the low variance filter and Pearson correlation exclude, is divided into two parts using the holdout method. The data is split into 80% for training data and the remaining 20% for testing data.

3.6. Decision Tree

During this stage, the classification process utilizing the Decision Tree method is carried out by employing the Decision Tree package provided in the big data analytics tool, KNIME. The model is constructed through a series of steps using the Decision Tree package. The initial step involves defining the Decision Tree package to invoke the Decision Tree library, which will be utilized for the creation of the classification model.

3.7. Evaluation Method

In this study, the overall classification performance is represented using a Confusion Matrix in tabular form. The rows of the Confusion Matrix display the predicted data, while the columns indicate the actual classes.TP represents the number of data correctly classified as positive, while FP represents the number of positive class data correctly classified as negative. TN represents the number of negative class data correctly classified as the negative class, and FN represents the number of negative class data correctly classified as the positive class. To evaluate the performance of the classification, accuracy, precision, and recall values are calculated. Precision and recall values are considered high if they reach or exceed a predetermined threshold. If the values reach 0.9, they can be categorized as high results. The accuracy is calculated using Equation 3, the precision is calculated using Equation 4, and the recall is calculated using Equation 5.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Recall} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

4. Results and Discussion

4.1. Low Variance Feature Selection

The low variance filter implements an optimization loop with a “best threshold” meta node to extract the optimal threshold value for feature removal. The low variance filter calculates the variance of each Bot and Brute Force feature and removes features with variance values below the best threshold, which is determined to be 0.016. Table 1 presents the selected features resulting from the low variance filter process.

Table 1. Low Variance Selection Feature Results

Bot	Web Attack Brute Force	SSH – Patator
Protocol	Protocol	Protocol
Flow Duration,	Flow Duration,	Flow Duration,
Fwd IAT Total,	Total Length of Fwd Packets	Fwd IAT Total,
Bwd IAT Total,	Fwd IAT Total,	Bwd IAT Total,
Fwd PSH Flags,	Bwd IAT Total,	Fwd PSH Flags,
SYN Flag Count,	Fwd PSH Flags,	SYN Flag Count,
PSH Flag Count,	Packet Length Mean	PSH Flag Count,
ACK Flag Count,	SYN Flag Count,	ACK Flag Count,
URG Flag Count dan	PSH Flag Count,	URG Flag Count dan
Init_Win_bytes_forward	ACK Flag Count, URG Flag Count dan Subflow Fwd Bytes Init_Win_bytes_Forward Init_Win_bytes_backward min_seg_size_forward	Init_Win_bytes_forward

4.2. Correlation Pearson Exclude

The second method applied for feature selection is Pearson correlation, which measures the correlation between features from the results of the low variance filter. According to Eid et al. [4], IDS datasets exhibit strong relationships between their features. Pearson correlation ranges from -1 to +1. If the correlation coefficient is -1, it indicates a perfect negative linear relationship between the two features being examined. If the correlation coefficient is +1, it indicates a perfect positive linear relationship between the two features. If the correlation coefficient is 0, it means there is no relationship between the two features under investigation. Table 2 presents the results of the strength of feature relationships for each attack type with thresholds ranging from 0.1 to 0.4. As the threshold is increased, the strength of the relationships decreases for the selected features.

Table 2. Feature Relationship for each Attack

Threshold	Selected Feature		
	Bot	Web Attack Brute Force	SSH - Patator
0.1	9	14	9
0.2	9	14	8
0.3	7	12	8
0.4	7	9	7

4.3. Data Partitioning

a. The dataset consists of BENIGN with Bot which 80% (3146 rows) allocated for data training and 20% (786 rows) allocated for testing data.
b. The dataset consists of BENIGN with Web Attack Brute Force which 80% (2412 rows) allocated for data training and 20% (612 baris data) allocated for testing data.
c. The dataset consists of BENIGN and Web Attack Brute Force classes, with 80% (2412 rows) allocated for training data and 20% (612 rows) for testing data.

4.4. Accuracy, Precision, and Recall

The classification performance is considered good with high accuracy and low computation time. Table 3 displays the accuracy, precision, and recall of the Decision Tree classification for each attack type based on selected features determined by the threshold. The highest accuracy for Bot is 99.5%, Web Attack Brute Force is 99.8%, and SSH - Patator is 99.7%, with no significant difference observed. Computation time is also calculated. Table 3 shows that precision for Bot is above 92% and falls into the high category. Web Attack Brute Force achieves precision values exceeding 99%, as does SSH - Patator. Recall values for all classes exceed 97% and fall into the high category. Overall, each class is correctly classified with high precision and recall, but optimizing classification time is crucial due to large and time-sensitive network traffic.

Table 3. Result of Accuracy, Precision, and Recall

Threshold	Accuracy			Precision			Recall		
	Bot	Web Attack Brute Force	SSH - Patator	Bot	Web Attack Brute Force	SSH - Patator	Bot	Web Attack Brute Force	SSH - Patator
0.1	98.9	99.8	99.7	98.5	98.4	99.7	99.5	100	99.8
0.2	94.9	99.7	99.6	99.5	99.3	99.3	97.2	99.7	99.8
0.3	98.1	99.5	99.4	98.5	100	99.7	99.5	99	99
0.4	99.5	99.5	99.7	97.4	99.3	99.6	99	99.7	99.7

4.5. Computation Time

Table 4 Computation Time (Seconds)

Threshold	Bot	Web Attack Brute Force	SSH - Patator
0.1	10	4	16
0.2	8	11	16
0.3	10	9	17
0.4	3	12	8

Table 4 presents the computation time required by the Decision Tree for classifying each attack using the selected features determined by the threshold compared to using all features. The computation time for Bot is 70 seconds, Web Attack Brute Force is 49 seconds, and SSH - Patator is 88 seconds. These results indicate a significant difference in computation time, with a drastic reduction in computation time for Bot becoming 3 seconds, Web Attack Brute Force becoming 4 seconds, and SSH - Patator becoming 8 seconds when using the selected features.

5. Conclusion

Features selection methods that combine low variance filter and correlation person exclude successfully make a classification model for Bot and Brute Force threats with good performance. It can reduce 83 features to 10 features selected for Bot, 15 features selected for Web Attack Brute Force, and 10 features selected for SSH - Potator. The highest precision value for Bot is 98.5% with threshold 0.1 and 0.3, Web Attack Brute Force's highest value is 100% with threshold 0.3, and *SSH* – Patator's highest value is 99.7% with threshold 0.3 and 0.4. The highest recall value for Bot is 99.5% with threshold 0.1 and 0.3, Web Attack Brute Force is 100% with threshold 0.1 and SSH - Patator is 99.8% with threshold 0.1 and 0.2

Reference

[1] Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

[2] Aksu, D. (2018). Intrusion Detection with Comparative Analysis of Supervised Learning Techniques and Fisher Score Feature Selection Algorithm. *Communications in Computer and Information Science,* 935 CCIS, 141–149. https://doi.org/10.1007/978-3-030-00840-6_16

[3] Amiri, F., Rezaei Yousefi, M., Lucas, C., Shakery, A., & Yazdani, N. (2011). Mutual information-based feature selection for intrusion detection systems. *Journal of Network and Computer Applications*, *34*(4), 1184–1199. https://doi.org/10.1016/j.jnca.2011.01.002

[4] Eid, H. F., Hassanien, A. E., Kim, T. hoon, & Banerjee, S. (2013). Linear Correlation-Based Feature Selection for Network Intrusion Detection Model. *Communications in Computer and Information Science*, *381 CCIS*, 240–248. https://doi.org/10.1007/978-3-642-40597-6_21

[5] Kundu, M. K., Mohapatra, D. P., Konar, A., & Chakraborty, A. (2014). Advanced computing, networking and informatics - Volume 1: Advanced computing and informatics proceedings of the second international conference on advanced computing, networking and informatics (ICACNI-2014). *Smart Innovation, Systems and Technologies*, *27* (VOL 1). https://doi.org/10.1007/978-3-319-07353-8

[6] Oreski, D., & Novosel, T. (2014). Comparison of Feature Selection Techniques in Knowledge Discovery Process. *TEM Journal*, *3*(4), 285–290. Retrieved from www.temjournal.com

[7] Saputra, F. A. (2018). Bot Detection in Network System Through Hybrid Low Variance Filter , Correlation Filter and Supervised Mining Process. *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, 112–117.

[8] Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). *Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization*. (Cic), 108–116. https://doi.org/10.5220/0006639801080116

[9] Sheena, Kumar, K., & Kumar, G. (2016). Analysis of Feature Selection Techniques: A Data Mining Approach. *International Conference on Engineering & Technology*, *4*(Icaet), 17–21. Retrieved from https://research.ijcaonline.org/icaet2016/number1/icaet024.pdf

[10] Silipo, R., Adae, I., Hart, A., & Berthold, M. (2014). Seven Techniques for Dimensionality Reduction. *Knime*, 1–21. Retrieved from https://www.knime.org/fails/knime_seventechniquesdatadimreduction.pdf